Capstone Project - 2

Supervised ML - Regression

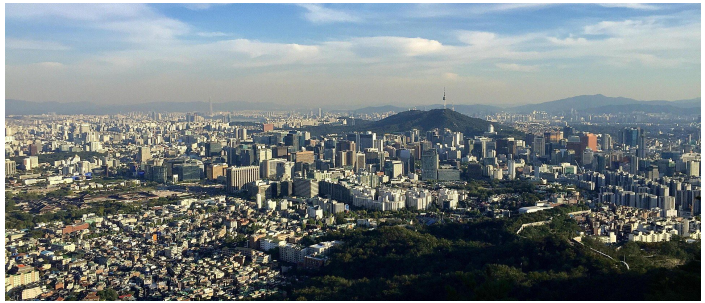
Topic - Bike Sharing Demand Prediction

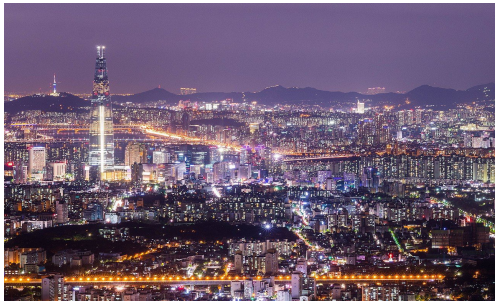
By - Avilash Srivastava

CONTENTS OF THE PRESENTATION

- Seoul City
- Problem Statement
- Data Summary
- Exploratory data analysis
- Data wrangling
- Machine Learning models
- Model Explanation
- Conclusion

Seoul City





The contents in the data belongs to the city called Seoul.

Seoul, officially the Seoul Special City, is the capital and largest metropolis of South Korea. Seoul has a population of 9.7 million people, and forms the heart of the Seoul Capital Area with the surrounding Incheon metropolis and Gyeonggi province. Considered to be a global city, Seoul was the world's 4th largest metropolitan economy in 2014 after Tokyo, New York City and Los Angeles.

Seoul has a humid continental climate influenced by the monsoons. Being in the extreme East Asia, the climate can be described as humid subtropical with great variation in temperature and precipitation throughout the year. Summers are hot and humid, with the East Asian monsoon taking place from June until September. August, the hottest month, has average high and low temperatures of 32.6 and 23.4 °C (91 and 74 °F) with higher temperatures possible. Heat index values can surpass 40 °C (104.0 °F) at the height of summer. Winters are quite short but usually cold to freezing with average January high and low temperatures of 1.5 and −5.9 °C (34.7 and 21.4 °F), and are generally much drier than summers, with an average of 24.9 days of snow annually. Sometimes, temperatures drop dramatically to below −10 °C (14 °F), and on some occasions as low as −15 °C (5 °F) in the mid winter period of January and February. Temperatures below −20 °C (−4 °F) have been recorded.

Problem Statement

- Bike rentals have became a popular service in recent years and it seems people are using it more often. With relatively cheaper rates and ease of pick up and drop at own convenience is what making this business excel. Mostly used by people having no personal vehicles and also to avoid congested public transport which follows its own time.
- Therefore, the business to strive and profit more, it has to be always ready and supply no. of bikes at different locations, to fulfil the demand. A pre planned set of bike count values can therefore, be a handy solution to meet all demands.

Data Summary

1. We have a variable 'date' which tells us the day on which some bikes were rented
2. We have a variable 'hour' which tells which hour of the day the bikes were rented.
3. We have some numerical type variables such as temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall which tells the environment conditions at that particular hour of the day.
4. We have some categorical variable such as seasons, is it a holiday and is it a functioning day or not.
5. And finally we have 'rented bike count' variable which we need to predict for new observations given the other variables.

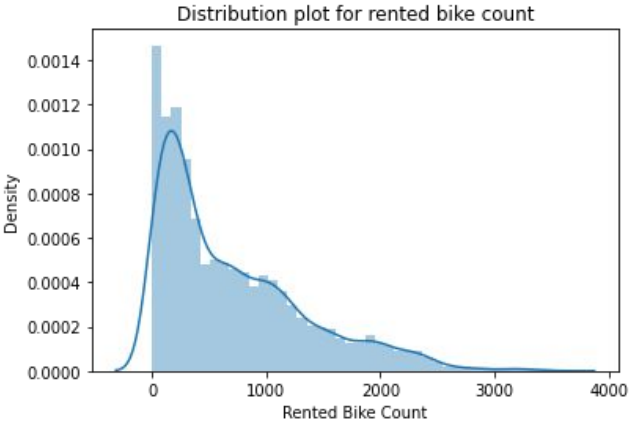
Total no of rows = 8760

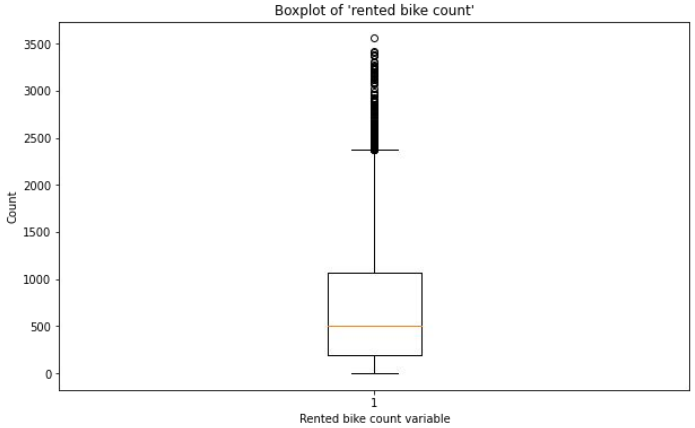
Total no of columns = 14

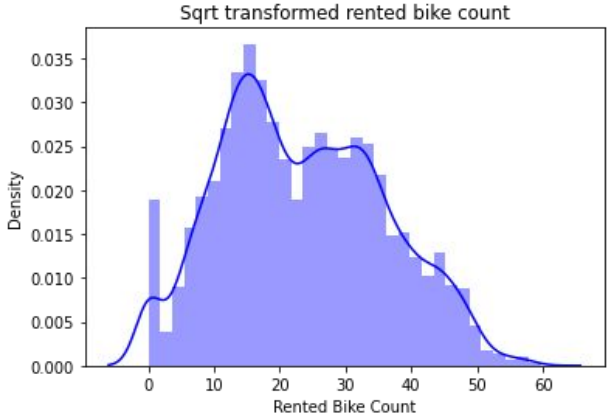
EDA - Univariate analysis

1. **"Rental Bike Count" - Dependent variable(dv)**
   - **Outliers above 2500**
   - **Was moderately skewed, positively**

```
Skewness: 1.153428  Skewness after transformation: 0.237362
Kurtosis: 0.853387  Kurtosis after transformation: -0.657201
```

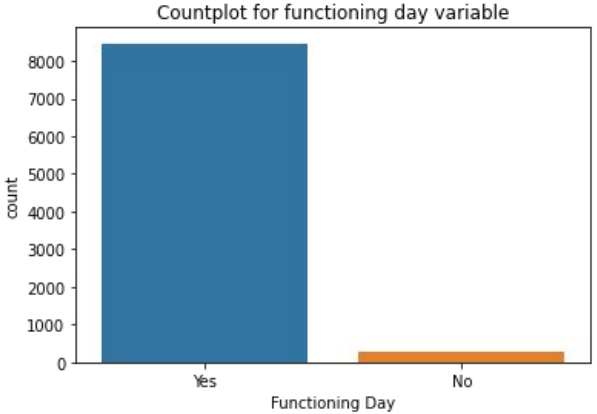


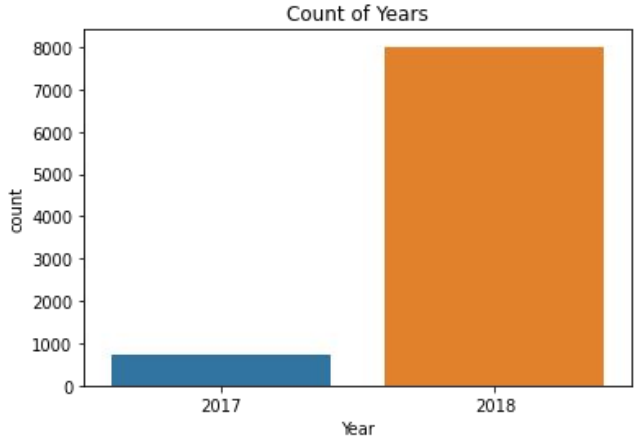


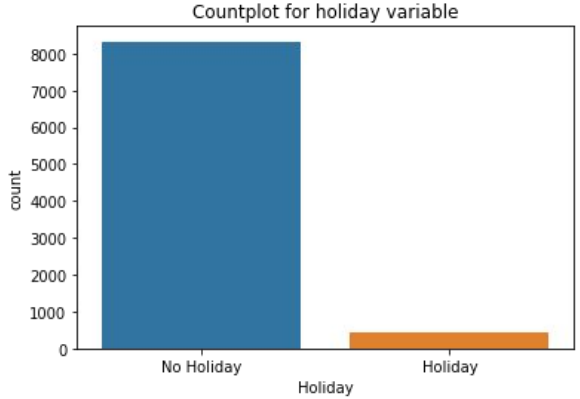


Univariate analysis - Categorical variables

- Functioning day and holiday had majority of one class around 97% and 95%.
- Majority of observations were from 2018.
- Decided to drop them.

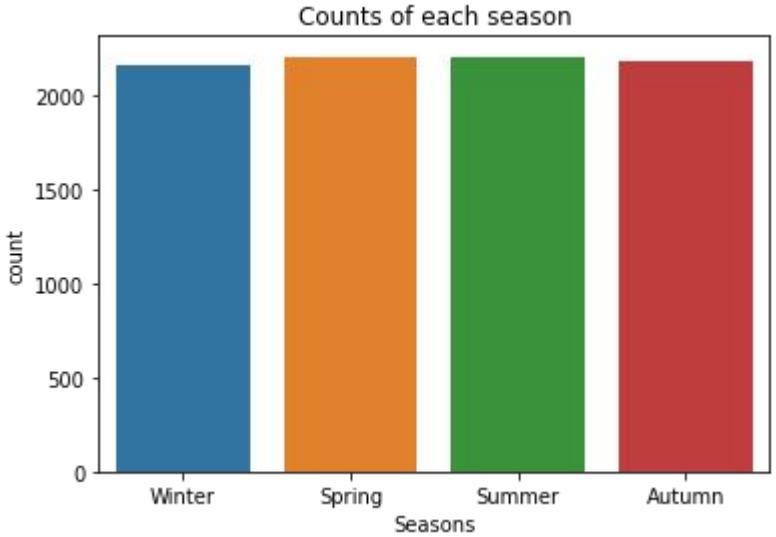


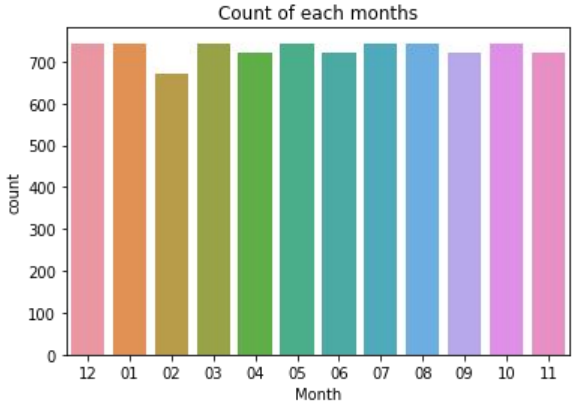


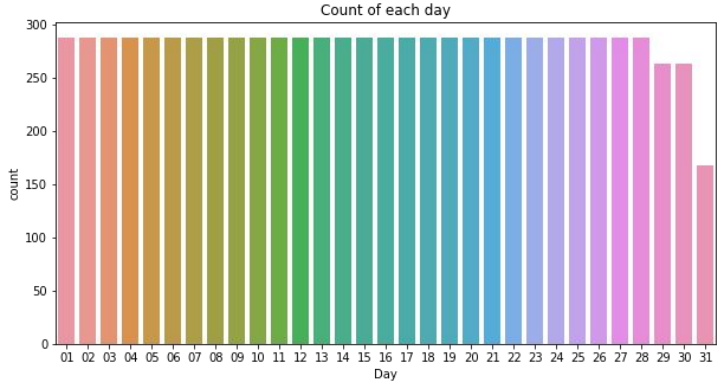


Univariate analysis- Categorical variables

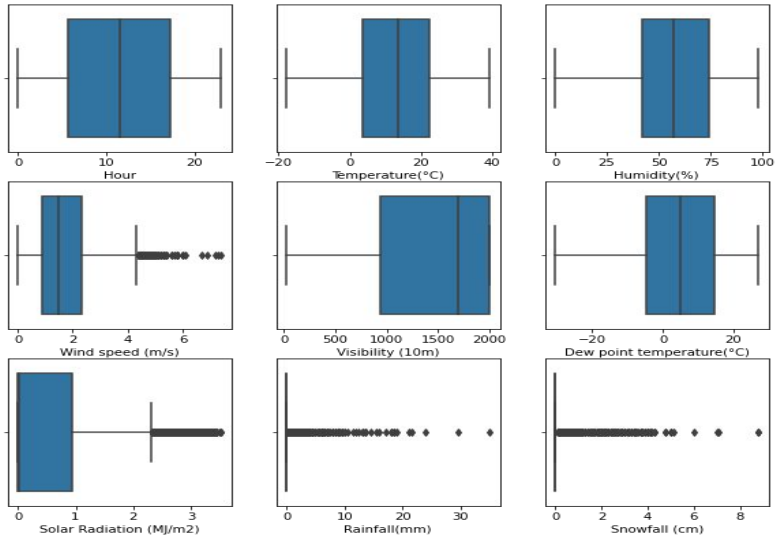
Seasons, day, month had almost equal no. of observations for each.

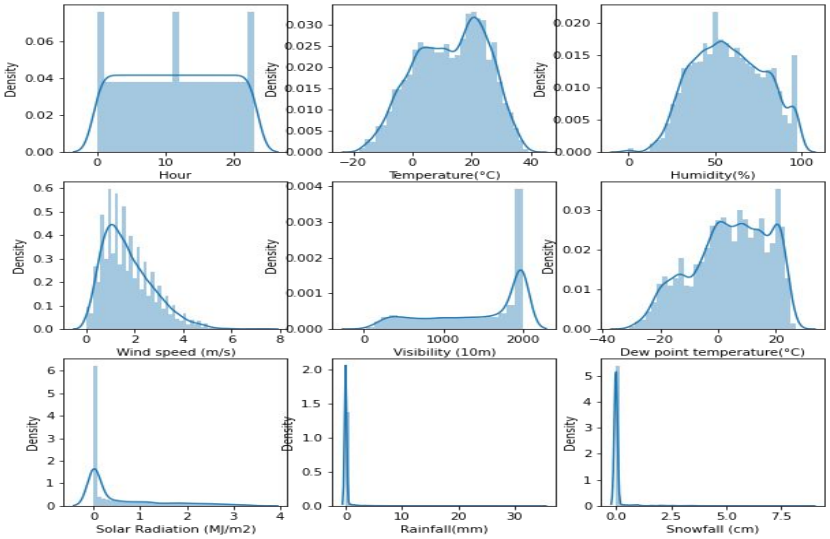






Univariate analysis - Numerical variables

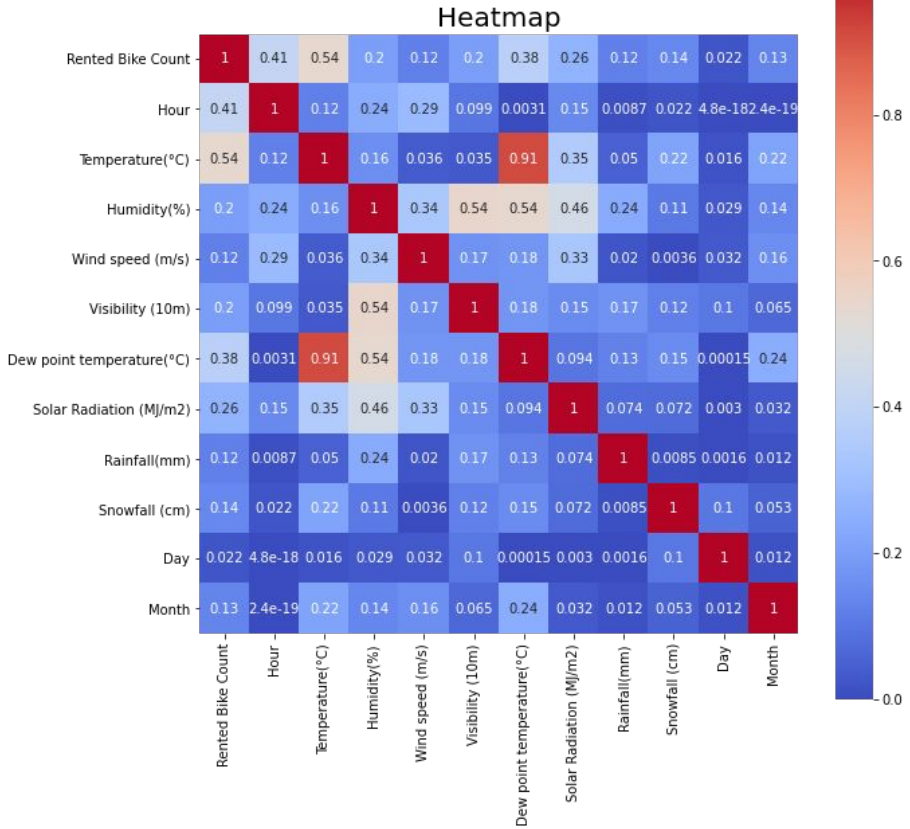




- Variables such as snowfall, rainfall had mostly zero values.
- Decided to drop them.

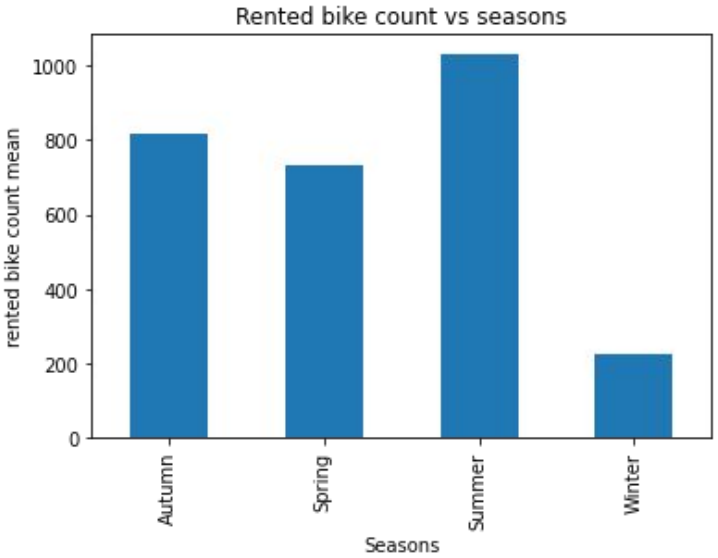
Multivariate analysis - Correlation

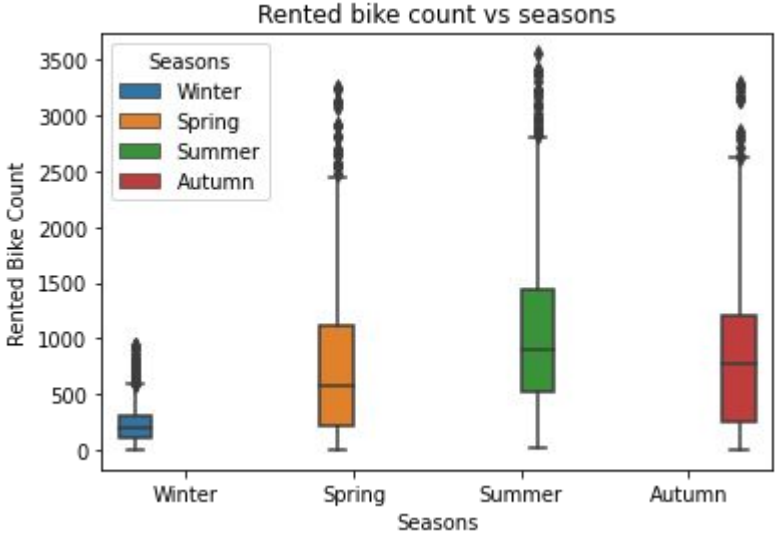
- Dew point temperature and Temperature were highly correlated.
- Linear regression assumes that independent variables must show some linear relationship with dependent variable.
- No such relationship seen here.
- Linear regression might not perform well.



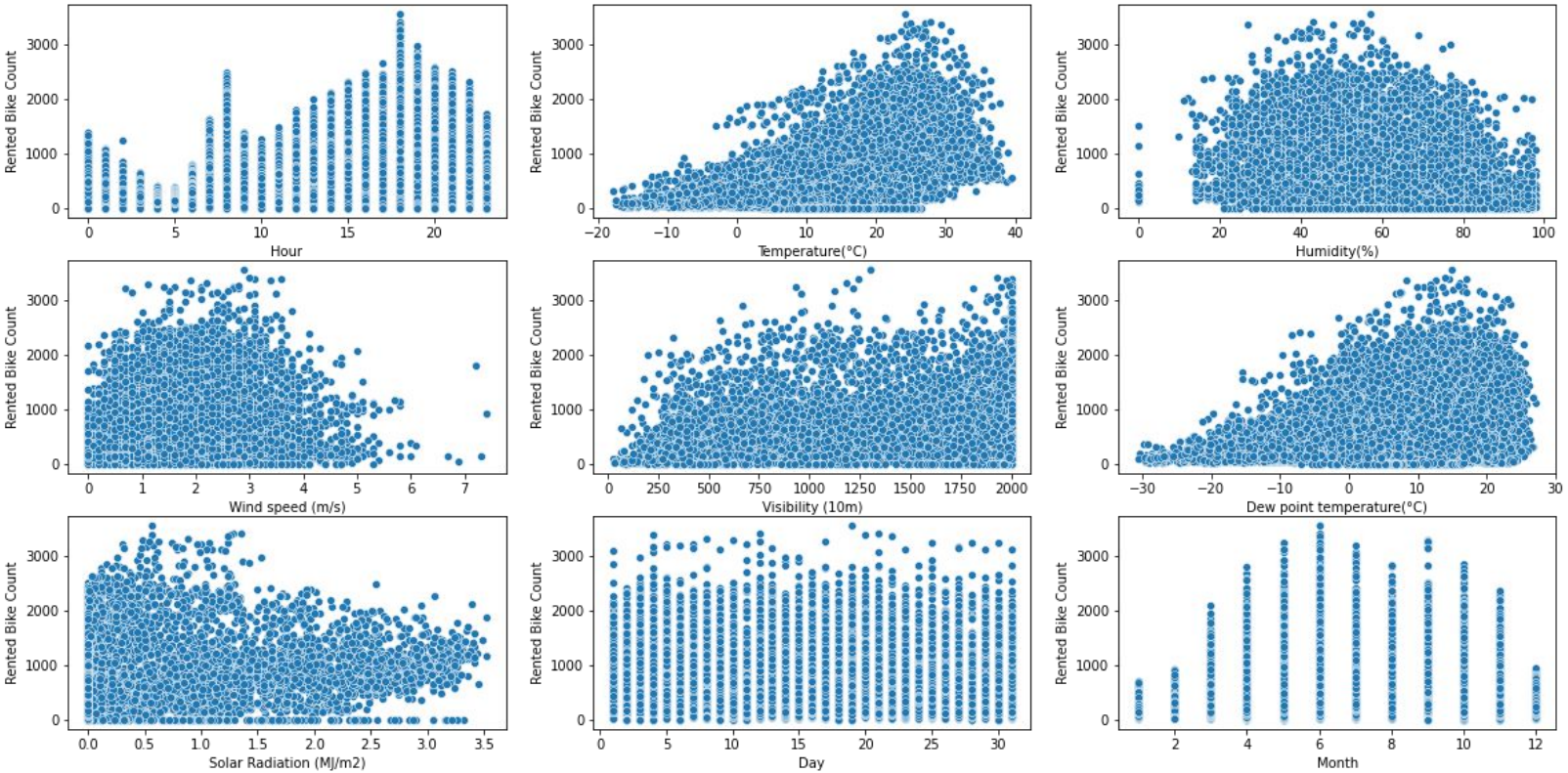
Multivariate analysis

Demand for bikes was high during summer compared to winter

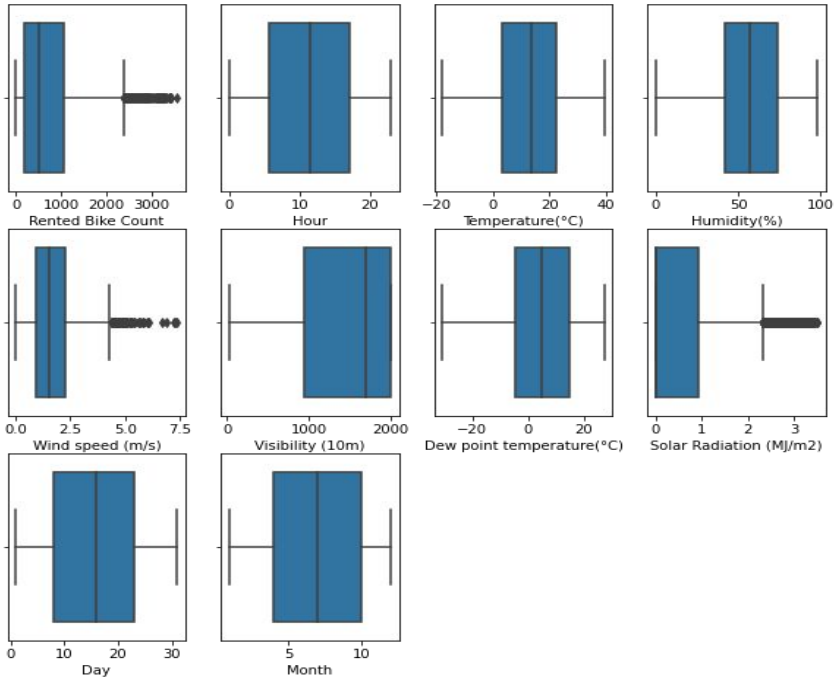




Multivariate analysis



Data Wrangling - missing values and outliers




```
The no. of missing values in each variable:
 Date                          0
Rented Bike Count              0
Hour                           0
Temperature(°C)                0
Humidity(%)                    0
Wind speed (m/s)               0
Visibility (10m)               0
Dew point temperature(°C)      0
Solar Radiation (MJ/m2)        0
Rainfall(mm)                   0
Snowfall (cm)                  0
Seasons                        0
Holiday                        0
Functioning Day                0
Day                            0
Month                          0
Year                           0
```

- No missing values
- Tackled outliers in rented bike count by applying transformation
- Windspeed and solar radiation had outliers, but they were not that far from maximum values.

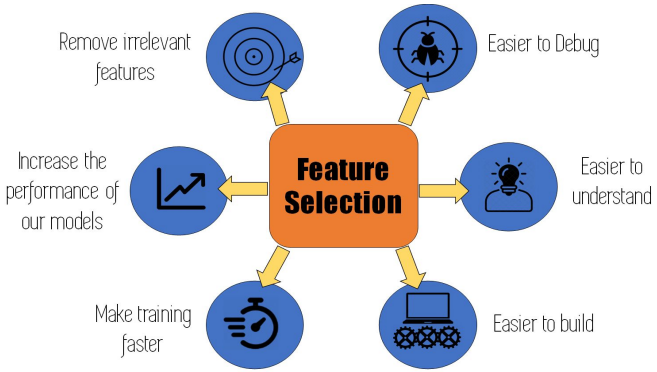
Data wrangling - feature selection

Removed columns :

- Functioning day, Holiday, Year - Had majority of one class
- Snowfall, Rainfall - Had mostly 0
- Dew point temperature - highly correlated with temperature

Encoding:

- One hot encoded Season



Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

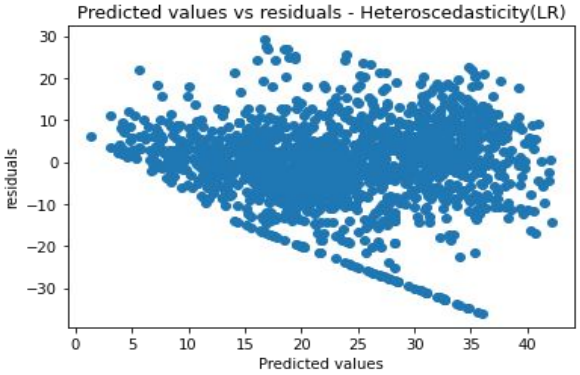
Machine learning Models

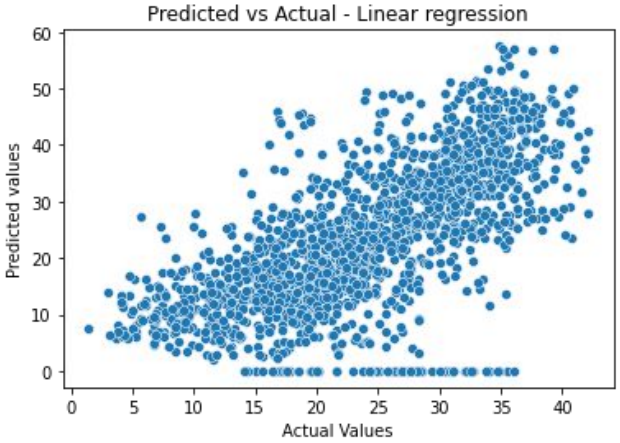
Three models were used : Linear regression, Random Forest, XGBoost. The evaluation results are:

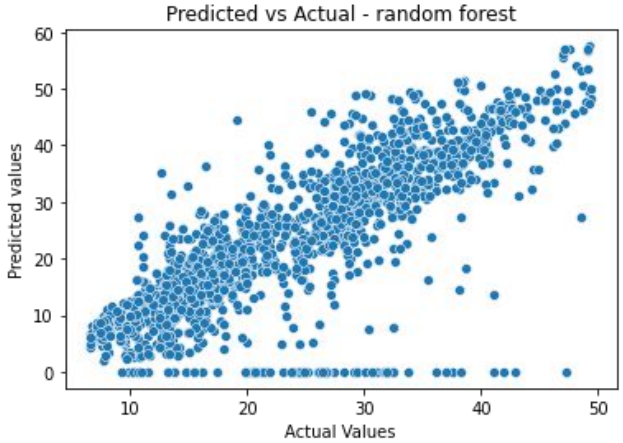
		Model	MAE	MSE	RMSE	R2_score	Adjusted R2	Comments
Training set	0	Linear regression	319.132	215574.806	464.300	0.481	0.48	Possible underfitting
	1	Random Forest - Before hyperparameter tuning	59.736	10809.661	103.970	0.974	0.97	Possible overfitting
	2	Random Forest - After hyperparameter tuning	195.709	95705.181	309.363	0.770	0.48	Good
	3	XGBoost - Before hyperparameter tuning	188.624	89520.703	299.200	0.784	0.78	Good
	4	XGBoost - After hyperparameter tuning	160.616	64595.842	254.157	0.844	0.84	Best of all model
Test set	0	Linear regression	323.570	219948.685	468.987	0.474	0.47	Possible underfitting
	1	Random Forest - Before hyperparameter tuning	171.947	88456.943	297.417	0.789	0.79	Possible overfitting
	2	Random Forest - After hyperparameter tuning	215.403	120972.279	347.811	0.711	0.71	Good
	3	XGBoost - Before hyperparameter tuning	204.973	107625.473	328.063	0.743	0.74	Good
	4	XGBoost - After hyperparameter tuning	184.669	82918.390	287.956	0.802	0.80	Best of all model

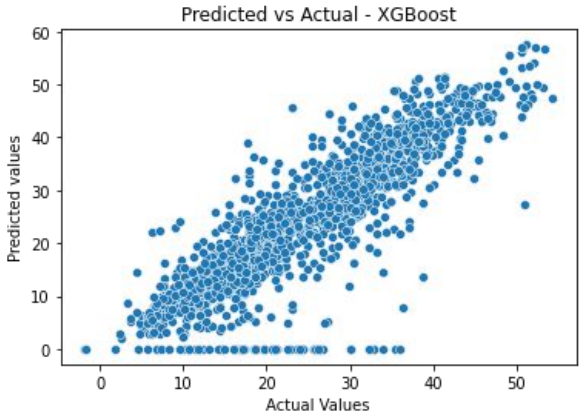
Model selection and validation

- Linear regression showed underfitting, heteroscedasticity
- Random forest overfitted initially
- Random forest got better after hyperparameter tuning
- XGBoost was good overall.









Best Model

- XGBoost outperformed all. After hyperparameter tuning the scores of evaluation metrics were:

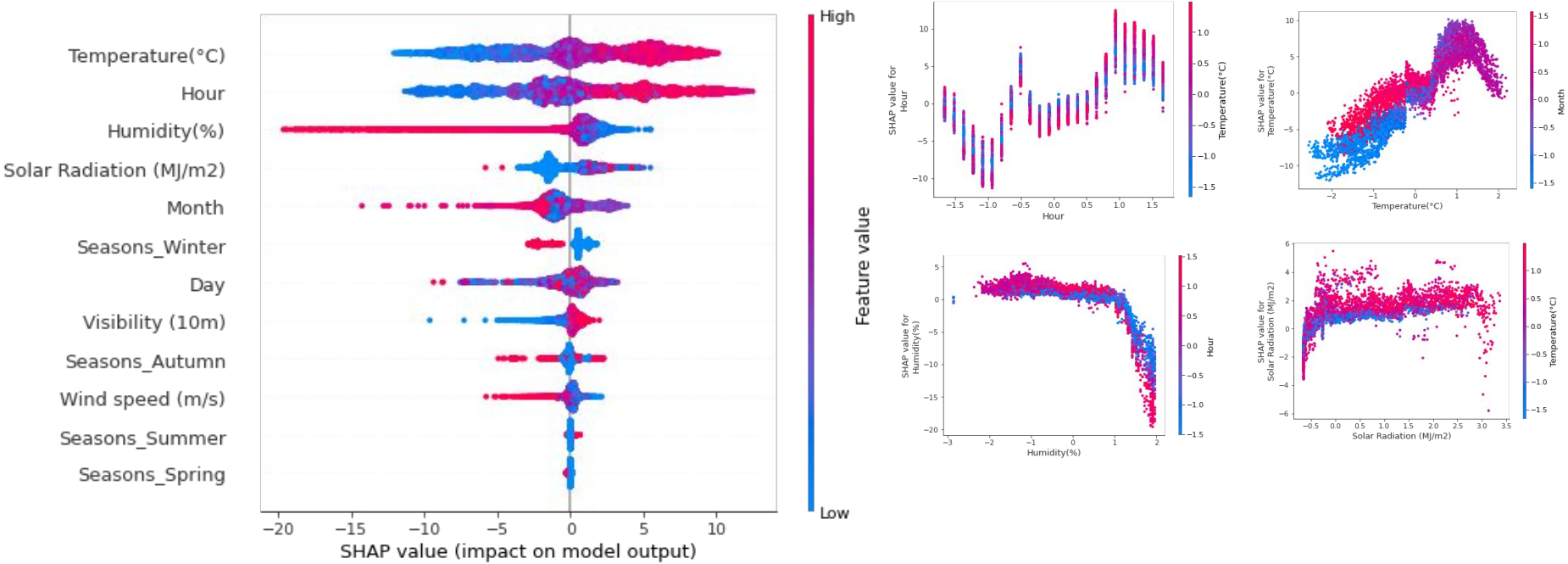
```
The evaluation metric values for training set - XGBoost
after hyperparameter tuning:
The MAE of training set =  160.6157715654452
The MSE of training set =  64595.84199944875
The RMSE of training set =  254.15712069396903
The R2_score of training set =  0.8444384642527141
The Adjusted R2_score of training set =  0.8441715967146202

The evaluation metric values for test set - XGBoost
after hyperparameter tuning:
The MAE of test set =  184.66897200803234
The MSE of test set =  82918.38953984065
The RMSE of test set =  287.9555339628684
The R2_score of test set =  0.8018783410327597
The Adjusted R2_score of test set =  0.800511199050237
```

The best hyperparameters were:

- Gamma = 0
- Learning rate = 0.28
- Max_depth = 4
- Min_child_weight = 1
- N_estimators = 46
- Subsamples = 0.5

Model Explanation

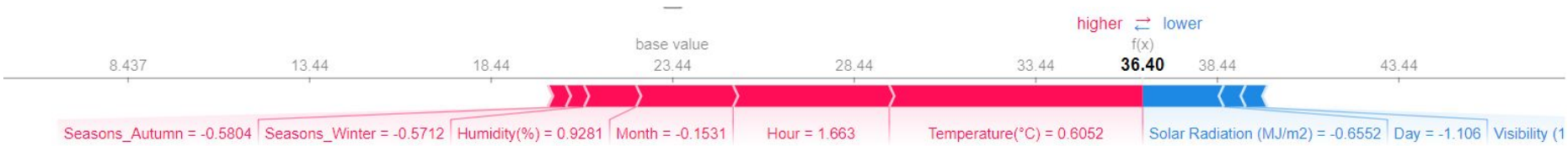
The most important features were Temperature, Hour, Humidity, Month.



Model Explanation - each observation

5th observation

100th observation



Conclusion

- We saw underfitting scenario in Linear regression and overfitting in Random Forest but the best performance was given by the XGBoost model.
- One of the challenges faced was to tune the hyperparameters, and find the best values which gave a better model.
- We also implemented shap techniques to understand the working of our XGBoost model and found out:
    1. Hour of the day had the most impact on predicting values. Demand was high during evening and night hours.
    2. Temperature was the second most important feature. Demand for bikes was higher when temperature was high.
    3. Demand was high for lower values of windspeed and solar radiation.
    4. Demand was less in winters as compared to other seasons.
- As this data is time dependent, the values for variables like temperature, windspeed, solar radiation etc., will not always be consistent. Therefore, there will be scenarios where the model might not perform well. As Machine learning is an exponentially evolving field, we will have to be prepared for all contingencies and also keep checking our model from time to time.